STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:4428–4440
Published online 14 March 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2865

An investigation of two multivariate permutation methods for controlling the false discovery proportion[‡]

Edward L. Korn[1,*,†], Ming-Chung Li[2], Lisa M. McShane[1] and Richard Simon[1]

[1]*Biometric Research Branch, National Cancer Institute, EPN-8129, Bethesda, MD 20892-7434, U.S.A.*
[2]*EMMES Corporation, Rockville, MD 20850, U.S.A.*

SUMMARY

Identifying genes that are differentially expressed between classes of samples is an important objective of many microarray experiments. Because of the thousands of genes typically considered, there is a tension between identifying as many of the truly differentially expressed genes as possible, but not too many genes that are not really differentially expressed (false discoveries). Controlling the proportion of identified genes that are false discoveries, the false discovery proportion (FDP), is a goal of interest. In this paper, two multivariate permutation methods are investigated for controlling the FDP. One is based on a multivariate permutation testing (MPT) method that probabilistically controls the number of false discoveries, and the other is based on the Significance Analysis of Microarrays (SAM) procedure that provides an estimate of the FDP. Both methods account for the correlations among the genes. We find the ability of the methods to control the proportion of false discoveries varies substantially depending on the implementation characteristics. For example, for both methods one can proceed from the most significant gene to the least significant gene until the estimated FDP is just above the targeted level ('top-down' approach), or from the least significant gene to the most significant gene until the estimated FDP is just below the targeted level ('bottom-up' approach). We find that the top-down MPT-based method probabilistically controls the FDP, whereas our implementation of the top-down SAM-based method does not. Bottom-up MPT-based or SAM-based methods can result in poor control of the FDP. Published in 2007 by John Wiley & Sons, Ltd.

KEY WORDS: false discovery rate; FDP; FDR; microarrays; multiple comparisons

1. INTRODUCTION

Analysis of gene expression profiles can involve tens of thousands of genes. To recognize the signal amid the noise leads to a multiple comparisons problem: when examining very many statistics,

*Correspondence to: Edward L. Korn, Biometric Research Branch, National Cancer Institute, EPN-8129, Bethesda, MD 20892-7434, U.S.A.
†E-mail: korne@ctep.nci.nih.gov

Received 26 May 2006
Published in 2007 by John Wiley & Sons, Ltd. *Accepted 26 January 2007*

some will appear large and interesting even when there is nothing truly happening. In this paper we will focus on identifying genes that are differentially expressed between two classes of expression profiles, e.g. microarray expression values obtained from normal tissue *versus* tumour biopsies. Such gene identification is an important goal of many microarray investigations [1]. The idea is to find as many genes that are truly differentially expressed while controlling the number of ('null') genes that are identified that are not truly differentially expressed (false positives or false discoveries). There is a trade-off involved in procedures for identifying differentially expressed genes: the more stringent the procedure is in keeping the number of false discoveries low, the less sensitivity there will be to detect truly differentially expressed genes. For example, suppose one uses a Bonferroni procedure and identifies genes from a set of 10 000 when their p-value from a two-sample univariate statistical test is less than $0.05/10\,000$. This procedure will result in one or more false discoveries less than 5 per cent of the time. On the other hand, suppose one used a procedure that did not control for multiple comparisons and identified all genes whose p-values were less than 0.05. Then, one could expect up to 500 ($=0.05\times 10\,000$) false discoveries. The trade-off is that the Bonferroni procedure will identify many fewer truly differentially expressed genes than the latter procedure.

A compromise between insisting on no false discoveries and making no adjustment for multiple comparisons is to allow for some false discoveries, but not too many. For example, suppose we identified all genes whose p-values were less than 0.001. With 10 000 genes, one would expect 10 false discoveries on average using this procedure if all the genes were null. Therefore, since some of the 10 000 genes are expected to be truly differentially expressed (i.e. non-null), the 10 false discoveries can be viewed as an upper bound for the expected number of false discoveries. Rather than controlling for the expected number of false discoveries, Benjamini and Hochberg [2] discussed controlling the expected false discovery proportion (FDP), which they called the false discovery rate (FDR). The FDP for a given set of identified genes is the proportion of genes in that set that are truly null. That is, $\mathrm{FDP}=V/D$, where D is the number of genes identified and V is the number of these genes that are null. (The FDP is defined to be zero when no genes are identified, i.e. when $D=0$.) Note that $\mathrm{FDR}=E(\mathrm{FDP})$ is a constant associated with the experimental design and analysis method, whereas the FDP is a random variable that will change from realization to realization of the data.

Since the FDP is a random variable, we would like to know in what way it is probabilistically controlled by an analysis method. Ideally, we would like to apply a procedure in a way that we can be $1-\alpha$ confident (e.g. $1-\alpha=80$ per cent confident) that the FDP (for the set of identified genes S) is less than γ (e.g. $\gamma=10$ per cent). In obvious notation, the condition is $P(\mathrm{FDP}(S)\leqslant\gamma)\geqslant 1-\alpha$. This provides more control than bounding the FDR [3], which has been the focus of much of the previous work in this area; see Ge *et al.* [4] and Li *et al.* [5] for extensive reviews.

Many methods have been proposed for finding differentially expressed genes. For this investigation, we wanted to examine non-parametric multivariate permutation methods which claim to provide control over false discoveries in a manner that could account for correlations among the genes and not require a large sample size (of arrays). We will not discuss analysis of variance models [6] or empirical Bayes methods [7–9] for expression data, which would be expected to be preferable when the number of arrays is very small and the distributional assumptions are satisfied [10].

One of the multivariate permutation procedures we consider is based on extensions of multivariate permutation tests (MPTs) that control for no false positives [11, 12] or a fixed number of false positives [3]. The other procedure we consider is an extension of Significance Analysis of

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2007; **26**:4428–4440
DOI: 10.1002/sim

Microarray (SAM) [13] that estimates the number of false positives for fixed cut-offs of functions of average class differences. SAM is a widely used procedure; the SAM paper [13] has been referenced more than 1750 times as of this date (ISI Web of Science). We describe the procedures in the next section and in Section 3 we evaluate their FDP-controlling properties. Section 4 presents an example involving genes that are differentially expressed in different types of breast cancer tumours. We end with a brief discussion of extensions to experimental designs other than the unpaired two-class comparison.

2. METHODS

Both the MPT-based methods and the SAM-based methods involve permuting the class labels to form new data sets on which various quantities are computed. Quantiles (e.g. the 90th percentile) of the permutation distribution of these quantities are then used to identify differentially expressed genes. We focus on the methods for unpaired two-class comparisons; other experimental designs are briefly considered in the Discussion.

2.1. MPT-based methods

MPT-based methods are based on the null hypothesis that the multivariate distribution of expression values for the null genes is the same in the two classes, with n arrays in one class and m arrays in the other class. We describe the method here; detailed justification is given by Korn *et al.* [3] and software is available [14]. First, calculate the two-sample t-statistic test for comparing the two classes for each gene i

$$t_i = \frac{\overline{x}_i - \overline{y}_i}{\hat{\sigma}_i \sqrt{1/n_i + 1/m_i}} \tag{1}$$

where $\overline{x}_i$ and $\overline{y}_i$ are the means of the gene expression values (suitably normalized) for each class, $\hat{\sigma}_i^2$ is the standard pooled variance estimator, and n_i and m_i are the available sample sizes in the two classes (which may be less than n and m because of missing data), all for gene i, where there are K genes in all. Let $p_1, p_2, \ldots, p_K$ be the normal-theory parametric p-values associated with these t-statistics and $p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(K)}$ be the ordered p-values. Consider B permutations of the class labels among the arrays. (Let $B \equiv (n+m)!/(n!m!)$, the number of possible permutations, if B is not too large a number. Otherwise, let $B = 1000$ random permutations.) For each permutation, calculate the two-sample t-statistic p-value for each of the genes using the permuted class labels and order the p-values from smallest to largest. Let $p_{(1),j} \leqslant p_{(2),j} \leqslant \cdots \leqslant p_{(K),j}$ denote these ordered p-values for the jth permutation, $j = 1, 2, \ldots, B$. Suppose we desire to identify a set of genes so that we could be $100(1-\alpha)$ per cent confident that there were at most u false discoveries in the set. This can be accomplished by identifying all genes whose p_i is less than the $(u+1)$st smallest p-value in 100α per cent of the permutations [3]. That is, one identifies all genes whose $p_i < \mathrm{MIN}_u(\alpha)$, where $\mathrm{MIN}_u(\alpha)$ is the 100αth percentile of $\{p_{(u+1),1}, p_{(u+1),2}, \ldots, p_{(u+1),B}\}$.

The MPT controls the number of false discoveries with $100(1-\alpha)$ per cent confidence regardless of the underlying distribution of the data; see the Discussion. However, $P(\mathrm{FDP} \leqslant \gamma)$ is not necessarily monotonic for nested gene lists $\{i \,|\, p_{(i)} \leqslant c\}$ for increasing c. That is, decreasing the significance level to identify differentially expressed genes does not mean decreasing the probability $P(\mathrm{FDP} \leqslant \gamma)$. In attempting to use the MPT to control the FDP, there are two ways to proceed, a 'top-down'

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2007; **26**:4428–4440
DOI: 10.1002/sim

Table I. Hypothetical data demonstrating identification of gene list of 13 genes controlling the FDP$\leqslant$10 per cent ($\gamma = 0.10$) with 80 per cent confidence.

Gene (i)	Allowable errors $\|i\gamma\|$	Observed p-value $p_{(i)}$	$\text{MIN}_0(0.2)$ (allows 0 errors)	$\text{MIN}_1(0.2)$ (allows 1 error)	...
(1)	0	0.0001	**0.0013**	0.0018	...
(2)	0	0.0001	**0.0013**	0.0018	...
(3)	0	0.0001	**0.0013**	0.0018	...
(4)	0	0.0002	**0.0013**	0.0018	...
(5)	0	0.0002	**0.0013**	0.0018	...
(6)	0	0.0005	**0.0013**	0.0018	...
(7)	0	0.0006	**0.0013**	0.0018	...
(8)	0	0.0007	**0.0013**	0.0018	...
(9)	0	0.0007	**0.0013**	0.0018	...
(10)	1	0.0012	0.0013	0.0018	...
(11)	1	0.0012	0.0013	**0.0018**	...
(12)	1	0.0016	0.0013	**0.0018**	...
(13)	1	0.0017	0.0013	**0.0018**	...
(14)	1	0.0025	0.0013	**0.0018**	...
(15)	1	0.0027	0.0013	**0.0018**	...
⋮		⋮	⋮	⋮	...

approach and a 'bottom-up' approach. Consider a list of the genes ordered by their observed p-values with the most significant gene (smallest p-value) at the top of the list (Table I). Along with a column of these p-values, the next columns contain the values $\text{MIN}_0(\alpha)$, $\text{MIN}_1(\alpha)$, etc., where $\alpha = 0.2$ in Table I. These column values represent the p-value cut-offs for identifying genes with allowance for 0 errors, 1 error, etc. In the top-down approach, we start at the top of the list and work down as long as the allowable number of false discoveries divided by the number of genes identified is less than γ: if genes (1), (2), ..., $(i-1)$ have already been identified, we identify gene (i) if either $p_{(i)} < \text{MIN}_{\|i\gamma\|}(\alpha)$ or $\|i\gamma\| > \|(i-1)\gamma\|$ ('automatic identification'), where $\|x\|$ denotes the greatest integer less than or equal to x. In the hypothetical example given in Table I, if we allowed 0 false discoveries (with 80 per cent confidence) then we would identify the first 11 genes; if we allowed one false discovery we would identify the first 13 genes. To control the FDP$<$10 per cent, proceeding from the top we sequentially can identify the first nine genes by comparing the observed p-value with the fourth column $\text{MIN}_0(0.2)$. (Bolded numbers identify which columns are to be used for comparison with the observed p-values.) The 10th gene is then an automatic identification. (Heuristically, automatic identification is used because if we are 80 per cent confident that there are no false discoveries in the first nine genes, then we are automatically 80 per cent confident that there are $\leqslant$1 false discoveries in the first 10 genes.) Genes 11–13 are identified because their p-values are less than $\text{MIN}_1(0.2)$ (bolded numbers), and the identification procedure stops.

In the bottom-up approach, we start at the bottom of the list of genes ordered by increasing p-values, and work up the list as long as the allowable number of false discoveries divided by the number of genes identified is less than γ: genes (1), (2), ..., (i) are identified where i is the largest index such that $p_{(i)} < \text{MIN}_{\|i\gamma\|}(\alpha)$. Table II gives a hypothetical example where the bottom-up approach identifies 12 genes and the top-down approach identifies four genes. The bottom-up approach will practically always identify at least as many genes as the top-down approach.

Table II. Hypothetical data demonstrating identification of gene list of four genes using top-down approach or 12 genes using bottom-up approach controlling the FDP$\leqslant$10 per cent with 80 per cent confidence.

Gene (i)	Allowable errors $\|i\gamma\|$	Observed p-value $p_{(i)}$	$\text{MIN}_0(0.2)$ (allows 0 errors)	$\text{MIN}_1(0.2)$ (allows 1 error)	...
(1)	0	0.0001	**0.0005**	0.0018	...
(2)	0	0.0001	**0.0005**	0.0018	...
(3)	0	0.0002	**0.0005**	0.0018	...
(4)	0	0.0004	**0.0005**	0.0018	...
(5)	0	0.0007	**0.0005**	0.0018	...
(6)	0	0.0008	**0.0005**	0.0018	...
(7)	0	0.0008	**0.0005**	0.0018	...
(8)	0	0.0010	**0.0005**	0.0018	...
(9)	0	0.0011	**0.0005**	0.0018	...
(10)	1	0.0012	0.0005	0.0018	...
(11)	1	0.0015	0.0005	**0.0018**	...
(12)	1	0.0017	0.0005	**0.0018**	...
(13)	1	0.0025	0.0005	**0.0018**	...
(14)	1	0.0027	0.0005	**0.0018**	...
(15)	1	0.0030	0.0005	**0.0018**	...
⋮		⋮	⋮	⋮	...

The one possible exception is when the top-down procedure stops at an automatic rejection and the bottom-up procedure identifies one less gene.

The top-down approach was used previously [3]; consideration of the bottom-up approach is considered for the first time in this paper.

2.2. SAM-based methods

The SAM procedure is quite complex. We present the main ideas here for the two-sample problem and refer the reader to Tusher *et al.* [13] and Chu *et al.* [15] for details. The score for gene i is the statistic

$$d_i = \frac{\overline{x}_i - \overline{y}_i}{s_0 + s_i} \tag{2}$$

where $s_i = \hat{\sigma}_i\sqrt{1/n_i + 1/m_i}$ is the standard error of the numerator. The score (2) is precisely a two-sample t-statistic (1) except for s_0, which is calculated to minimize the coefficient of variation of the d_i. In calculating s_0, SAM uses a grid of 100 cut-points to define windows of increasing values of s_i; see Chu *et al.* [15]. If there are $K\leqslant 100$ genes, then s_0 is undefined and we take it to be zero in what follows. We will also present results for a SAM-based method that sets $s_0, = 0$ ('SAM without s_0') for $K > 100$. The score statistics are ordered for the K genes: $d_{(1)}\leqslant d_{(2)}\leqslant\cdots\leqslant d_{(K)}$. The microarray data are permuted as in the MPT-based method, and the $d_{(i)}$ are calculated for each permuted data set, denoted $d_{(i),j}$ for the jth permuted data set. (The quantity s_0 is not recalculated for each permuted data set, but s_0 from the original data set is used throughout.) Let $\overline{d}_{(i)}$ be the mean of the $d_{(i),j}$ across the permuted data sets. For example, $\overline{d}_{(K)}$ is the mean across the permuted data sets of the largest scores. For a fixed 'tuning parameter' Δ, genes are identified as follows. Find the smallest i_1 such that $d_{(i_1)}$ and $\overline{d}_{(i_1)}$ are positive and such that $d_{(i_1)} - \overline{d}_{(i_1)} > \Delta$.

Identify genes corresponding to $d_{(i_1)}, d_{(i_1+1)}, \ldots, d_{(K)}$ as 'positive'.Similarly, find the largest i_2 such that $d_{(i_2)}$ and $\overline{d}_{(i_2)}$ are negative and such that $\overline{d}_{(i_2)} - d_{(i_2)} > \Delta$. Identify genes corresponding to $d_{(1)}, d_{(2)}, \ldots, d_{(i_2)}$ as 'negative'.

To estimate the number of false positives for a given Δ, SAM uses the following procedure. First, define $\text{cut}_{\text{up}}(\Delta) = d_{(i_1)}$ and $\text{cut}_{\text{low}}(\Delta) = d_{(i_2)}$. (The $\text{cut}_{\text{up}}(\Delta)$ ($\text{cut}_{\text{low}}(\Delta)$) remains undefined if there are no positive (negative) genes identified.) For each permuted data set, let c^* be the number of genes with $d_i \geqslant \text{cut}_{\text{up}}(\Delta)$ plus the number of genes with $d_i \leqslant \text{cut}_{\text{low}}(\Delta)$. Across the permutations, calculate the 90th percentile of the c^*. (The SAM software [15] also allows the use of the median of the c^*, and the original description of SAM [13] used the mean of the c^*.) Multiply this 90th percentile by an estimate $\hat{\pi}_0$ of the proportion of true null genes; see Chu *et al.* [15] for a description of $\hat{\pi}_0$. The product is taken as the estimate of the number of false positives. The FDR is estimated by this estimate of false discoveries divided by the number of identified genes. The number of identified genes, estimated number of false positives, and estimated FDR can be displayed for a grid of Δ values. The SAM software [15] chooses the grid to be 100 values; in our simulations we choose the 100 values to correspond to the 100 percentiles (first percentile, second percentile, etc.) of the $|d(i) - \overline{d}(i)|$.

Although no probability claims are made for the method in Tusher *et al.* [13], it is of interest to assess the performance of SAM for controlling the FDP by choosing a Δ so that the estimated FDR from SAM is $\leqslant \gamma$. Let $\Delta_1 \geqslant \Delta_2 \geqslant \cdots \geqslant \Delta_{100}$ be the grid values of Δ. A top-down method of gene identification chooses the genes identified with Δ_i, where i is the smallest index such that the estimated FDR from SAM associated with Δ_i is $\leqslant \gamma$ and the estimated FDR from SAM associated with Δ_{i+1} is $> \gamma$. A bottom-up method of gene identification chooses the genes identified with Δ_i, where i is the largest index such that the estimated FDR from SAM associated with Δ_i is $\leqslant \gamma$ and the estimated FDR from SAM associated with Δ_{i+1} is $> \gamma$. Note that the bottom-up method will always identify at least as many genes as the top-down method, because a smaller Δ identifies more genes.

The bottom-up approach is one of the approaches that has been suggested for SAM [16]; the top-down approach is considered for the first time in this paper.

3. RESULTS

We consider some of the properties of the MPT-based and SAM-based methods for controlling the FDP. For the simulations, we generally assume that: (a) there are the same number of observations in each group and no missing data ($n_i = m_i = n$); (b) there are 100, 1000, or 5000 genes ($K = 100$, 1000, or 5000); (c) $\gamma = 10$ per cent; and (d) the observations are normally distributed. In particular, the observations are normally distributed with, for gene i, the same variance for each group (denoted σ_i^2) and mean shift between the groups being μ_i. The σ_i^2 are sampled from a distribution that is $0.25 + X/6.67$, with X having a chi-squared distribution with 5 degrees of freedom. The 0.25 term is used so that the variances will not be unrealistically close to zero, and the 6.67 factor is used so that the distribution has mean 1. When in a simulation some genes are differentially expressed ($\mu_i \neq 0$), we will express the differential expression in terms of the effect size (μ_i/σ_i) and specify that the distribution of the σ_i^2 is the same for differentially expressed and non-differentially expressed genes. (With these specifications, the MPT-based simulation results do not depend on the distribution of the σ_i^2.) Unless otherwise specified, the correlation between the genes is taken to be 0. All simulations are based on 10 000 repetitions. The simulations considered are designed

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2007; **26**:4428–4440
DOI: 10.1002/sim

Table III. Simulated proportion of times true FDP is greater than 10 per cent using MPT-based methods (80 per cent confidence) or SAM-based methods (90th percentile) for two-class problem under global null hypothesis.

Top-down or bottom-up	SAM	MPT
Top-down	0.173	0.199
Bottom-up	0.173	0.199

Note: sample size $n = 30$ per group, $K = 100$ independently normally distributed genes.

to demonstrate, in settings as simple as possible, the properties that are noted heuristically for the methods.

The first property to note is that using the SAM-based 90th percentile method, the probability that the FDP is larger than the specified γ can be as large as approximately 20 per cent, not the 10 per cent one might expect because the 90th percentile is used. This doubling of what one might think is the error rate is because of the one-sided nature of SAM [17]. The simulation results given in Table III under the global null hypothesis and with $n = 30$ and $K = 100$ genes demonstrate this; the SAM-based method yields a FDP$>\gamma$ ($= 10$ per cent) in 17.3 per cent of the simulations. In all the simulations that follow we will therefore compare the 90th percentile SAM-based method with the 80 per cent confidence MPT-based method.

As noted in Section 2, the SAM software uses a grid of 100 Δ values. If, as typically is the case, there are more than 100 genes, there is the possibility of using a larger number of Δ values in the grid. Increasing the number of Δ's in the grid (by adding more Δ's to the existing Δ's) can increase the number of genes identified by the bottom-up approach; for the top-down approach it can result in either an increase or decrease in the number of identified genes. Choosing $K\Delta$'s as the maximum number is reasonable because choosing more will not result in any differences in the genes identified. It is not clear how choosing 100 *versus* $K\Delta$'s will affect the properties of the SAM-based method for controlling the FDP. Table IV presents simulation results evaluating the effect on the SAM-based method of using 100 Δ's *versus* 1000 Δ's when $K = 1000$ and 12 genes are non-null and the rest are null. With 1000 Δ's the average number of non-null genes identified is 12.0. In fact, these 12 genes were identified in $\geqslant$9994 of the 10 000 simulated data sets by all of the SAM-based methods with 1000 Δ's. With 100 Δ's, only 10.1 of the 12 non-null genes were identified on average, and in <1 per cent of the simulated data sets all 12 non-null genes were identified. In all the simulations that follow, we set the number of Δ's equal to the number of genes, K. We note in passing that the MPT-based method also identified all 12 non-null genes for all the simulated data sets for this simulation.

As we mentioned above, the SAM-based methods would be expected to result in the FDP$>\gamma$ no more than 20 per cent of the time when the 90th percentile method is used. Although we demonstrated this in Table III under the global null hypothesis, this does not have to be the case when there are some non-null genes as the following heuristic argument shows. Suppose out of the K genes, K_+ are highly differentially expressed and K_0 are null. The SAM-based method will practically always identify at least the K_+ highly differentially expressed genes. Now, consider the distribution of the score for the gene among the K_0 null genes that is observed to have the most differential expression. This distribution will be the most extreme distribution from K_0 null distributions. However, it will be compared to the reference permutation distribution of the $(K_+ + 1)$ most extreme distribution out of K null distributions. This permutation distribution will

Table IV. Simulated proportion of times true FDP is greater than 10 per cent and average number of non-null genes identified using MPT-based methods (80 per cent confidence) or SAM-based methods (90th percentile) for two-class problem with 12 non-null genes.

	SAM (SAM without s_0 in parentheses)					
	100 Δ's		1000 Δ's		MPT	
Top-down or bottom-up	FDP>10 per cent	Number of non-null identified	FDP>10 per cent	Number of non-null identified	FDP>10 per cent	Number of non-null identified
Top-down	0.000 (0.000)	10.1 (10.1)	0.223 (0.229)	12.0 (12.0)	0.198	12.0
Bottom-up	0.000 (0.000)	10.1 (10.1)	0.223 (0.229)	12.0 (12.0)	0.198	12.0

Note: sample size $n = 30$ per group, $K = 1000$ independently normally distributed genes, six genes with effect size of 2 and six genes with effect size of -2.

Table V. Simulated proportion of times true FDP is greater than 10 per cent using MPT-based methods (80 per cent confidence) or SAM-based methods (90th percentile) for two-class problem with eight non-null genes.

Top-down or bottom-up	Number of genes	SAM (SAM without s_0 in in parentheses)	MPT
Top-down	100	0.361	0.188
Bottom-up	100	0.432	0.276
Top-down	1000	0.204 (0.212)	0.202
Bottom-up	1000	0.306 (0.312)	0.299

Note: sample size $n = 30$ per group, $K = 100$ or 1000 independently normally distributed genes, four genes with mean shift of 2 and four genes with mean shift of -2.

be less extreme than the observed distribution. For example, the ninth most extreme distribution out of 100 null distributions is less extreme than the most extreme distribution out of 92 null distributions. The end result is that the SAM-based method will reject null genes too often. This is demonstrated in Table V where the error rates for the SAM-based method are 36.1 and 43.2 per cent instead of being $\leqslant 20$ per cent. With $K = 1000$, the SAM-based method performs better than with $K = 100$ (see Table V): using a reference distribution of the ninth largest out of 1000 for an actual distribution of the largest out of 992 is not as large a problem as using a reference distribution of the ninth largest out of 100 for an actual distribution of the largest out of 92. In fact, the top-down SAM-based method has an acceptable error rate for this situation when $K = 1000$. However, when correlation is added to the genes, neither the top-down or bottom-up approach of the SAM-based method has acceptable error rates with $K = 1000$ or 5000 (Table VI). Changing the sample size from $n = 30$ per group to $n = 50$ per group or $n = 10$ per group, or making the sample size 30 in one group and 50 in the other does not improve the behaviour of the SAM-based method, with the results being almost the same as the results given in Table VI (data not shown).

Table VI. Simulated proportion of times true FDP is greater than 10 per cent using MPT-based methods (80 per cent confidence) or SAM-based methods (90th percentile) for two-class problem with eight non-null genes.

Top-down or bottom-up	Number of genes	SAM (SAM without s_0 in in parentheses)	MPT
Non-null genes all in one cluster			
Top-down	1000	0.267 (0.268)	0.198
Bottom-up	1000	0.287 (0.289)	0.271
Top-down	5000	0.244 (0.238)	0.195
Bottom-up	5000	0.271 (0.268)	0.277
Non-null genes all in different clusters			
Top-down	1000	0.265 (0.268)	0.198
Bottom-up	1000	0.285 (0.289)	0.271
Top-down	5000	0.244 (0.238)	0.195
Bottom-up	5000	0.272 (0.268)	0.278

Note: sample size $n = 30$ per group, $K = 1000$ or 5000, clusters of 50 genes with correlation $= 0.5$ within cluster, normally distributed genes, four genes with effect size of 2 and four genes with effect size of -2.

Tables V and VI demonstrate a potential problem with the bottom-up approach for both the SAM-based and MPT-based methods and the top-down approach for the SAM-based method; these approaches can violate the condition FDP$\leqslant$10 per cent more than 20 per cent of the time. This simulation is meant to be a difficult test for the methods, for if even one null gene is identified in addition to the eight non-null genes, the FDP will be greater than 10 per cent. A less difficult test is offered by the scenario in which 100 genes are non-null (Table VII), as the methods will still satisfy FDP$\leqslant$10 per cent even with 10 null genes identified. The SAM-based method depends on the signs (positive or negative) of the non-null genes (the MPT-based method does not), which is why two scenarios are considered in Table VII. The condition FDP$\leqslant$10 per cent is violated less than 20 per cent of the time in the Table VII simulations.

With smaller sample sizes than $n = 30$ per group, the advantages of pooling variability information (as SAM does) should be greater. In these situations, one might also consider pooling variability information for the MPT-based method also, e.g. using the method of Wright and Simon [18].

Although the focus of this paper is on controlling the FDP, the simulations can also be used to evaluate the FDR by averaging the FDPs across the simulated data sets. A special case is the global null hypothesis (Table III), for which the FDP can only be zero or one. In this special case, the simulated FDR is the same as the simulated proportion of times the true FDP is greater

Table VII. Simulated proportion of times true FDP is greater than 10 per cent and average number of non-null genes identified using MPT-based methods (80 per cent confidence) or SAM-based methods (90th percentile) for two-class problem with 100 non-null genes.

	SAM (SAM without s_0 in parentheses)					
	50 genes shift $=+\sigma_i$, 50 genes shift $=-\sigma_i$		90 genes shift $=+\sigma_i$, 10 genes shift $=-\sigma_i$		MPT	
					100 genes $\lvert\text{shift}\rvert=\sigma_i$	
Top-down or bottom-up	FDP>10 per cent	Number of non-null identified	FDP>10 per cent	Number of non-null identified	FDP>10 per cent	Number of non-null identified
Top-down	0.128 (0.149)	84.5 (85.6)	0.096 (0.114)	85.9 (86.6)	0.122	85.5
Bottom-up	0.135 (0.159)	84.8 (85.7)	0.103 (0.123)	86.0 (86.8)	0.129	85.7

Note: sample size $n=30$ per group, $K=1000$ independently normally distributed genes.

than 10 per cent, i.e. the proportions given in Table III. For the simulations in Tables IV–VII, the simulated FDR is always less than 10 per cent (data not shown).

4. EXAMPLE

Hedenfalk *et al.* [19] analysed cDNA microarray profiles from breast cancer tumours from patients who had a family history of breast or ovarian cancer and whose tumours had BRCA1 mutations (seven patients) or BRCA2 mutations (eight tumours from seven patients), as well as tumours from seven patients with sporadic cases of breast cancer. We compare the BRCA1 tumours ($n=7$) to the non-BRCA1 tumours ($n=15$), and the BRCA2 tumour ($n=8$) to the non-BRCA2 tumours ($n=14$) for the 3226 genes that met quality control standards. The data are available online (http://linus.nci.nih.gov/~brb/book.html) and additionally described elsewhere [20]. We use a target FDP$\leqslant$10 per cent and 80 per cent confidence level of the MPT-based methods and the 90th percentile for the SAM-based methods with 3226 Δ's. Because these permutation-based methods can yield variable results depending upon the random permutations, we performed each method 11 times (with 1000 permutations each time) and report here the median resulting number of identified genes for each method.

The methods identified roughly similar numbers of genes for both comparisons (Table VIII). As both the MPT-based method and the SAM (without s_0)-based method are using the same ordering of the genes (based on the two-sample t-statistic), the genes identified by these methods are subsets of each other. For example, the list of the 56 genes identified by top-down MPT-based method for the BRCA1 comparison is a subset of the list of 63 genes identified by bottom-up MPT-based method, which in turn is a subset the list of the 68 genes identified by the top-down or bottom-up SAM (without s_0)-based method. There is an overlap with the genes identified by the SAM-based method, but the overlap is not complete. For example, of the 72 genes identified by SAM, 59 of these were also identified by SAM without s_0.

Published in 2007 by John Wiley & Sons, Ltd. *Statist. Med.* 2007; **26**:4428–4440
DOI: 10.1002/sim

Table VIII. Number of genes identified for Hedenfalk *et al.* [19] data for two comparisons using various multivariate permutation-based methods.

Method	Top-down or bottom-up	BRCA1 *versus* non-BRCA1	BRCA2 *versus* non-BRCA2
MPT (80 per cent confidence)	Top-down	56	67
	Bottom-up	63	82
SAM without s_0 (90th percentile, 3226 Δ's)	Top-down	68	52
	Bottom-up	68	52
SAM (90th percentile, 3226 Δ's)	Top-down	72	70
	Bottom-up	72	77

Based on the simulations presented in Tables V and VI, we would recommend using the gene lists identified by the top-down MPT-based method because this is the only method that guarantees the putative confidence level.

5. DISCUSSION

In our implementation of the MPT and SAM-based methods for controlling the FDP we have used parametric p-values from two-sample t-statistics to rank the genes for MPT-based methods and parametric-type scores to rank the genes for the SAM-based methods. However, since the analyses are based on the multivariate permutation distribution and the p-values and scores are solely used to rank the genes in the observed and permuted data sets, the inference is, in fact, non-parametric provided that the multivariate distribution of the null gene values is the same in the two classes. Nevertheless, one could consider using a non-parametric statistic to rank the genes, for example, the Wilcoxon rank-sum test for an unpaired two-class comparison [21]. Although we believe that generally the use of parametric statistics rather than non-parametric statistics will lead to more non-null genes being identified, the choice of what is the best ranking statistic to use in which applications is an area of further research.

It is straightforward to apply the MPT-based methods to designs other than the unpaired two-class comparisons thus far discussed. One need only to apply an appropriate statistical method that yields a p-value for each gene and permute the labels of the microarray profiles consistently with the experimental design. For example, for a paired two-class comparison, one would use a paired t-statistic instead of the unpaired t-statistic, and one would permute the class labels within pairs (2^n possible permutations for n pairs of samples). For a regression problem where each microarray is associated with a single continuous covariate x, one could use a linear regression coefficient for the ith gene divided by its standard error instead of the unpaired t-statistic, and permute all the class labels among the x's ($n!$ possible permutations for n samples). For a C class problem with $C = 3$, one would use a standard F statistic $= [SS_b/(3-1)]/[SS_w/(n-3)]$ from the analysis of variance, where SS_b is the between-class sum of squares and SS_w is the within-class sum of squares for the ith gene. The microarray labels are permuted among the three classes ($(n_1+n_2+n_3)!/(n_1!n_2!n_3!)$ possible permutations with n_c samples in the cth class, $n = n_1 + n_2 + n_3$). Finally, one may not want to use p-values to rank the genes in more complex problems [22].

With the SAM-based methods, the method of performing the permutations is identical to the method for the MPT-based methods. However, the choice of the score for ranking the genes is

not canonical even for simple problems; see Chu *et al.* [15] for their recommended score choices for the paired two-class comparison and linear regression, and Tusher *et al.* [13] for a recommend score for a K class problem. We note that for any reasonable score for the K class problem with $K \geqslant 3$, the score will be unidirectional. That is, larger values of the score will represent larger class differences and smaller values will represent smaller class differences. This is distinct from the two-class problems where large or small values of the score represent class differences, and in distinction to the regression problem where large or small values of the score represent an association between the gene expression and covariate. Because of this distinction, the properties of the SAM-based methods for FDP control may be quite different for the K class problem than for the other types of experimental design. Therefore, the properties and results of the SAM-based methods discussed in this paper for the two-class comparison may not be relevant for the K class problem. This is not an issue for the MPT-based methods because they are always based on two-sided p-values.

In summary, if one desires to control with a specified confidence that the FDP is less than a specified value, then we recommend using the top-down MPT-based method.

ACKNOWLEDGEMENTS

This study utilized the high-performance computational capabilities of the Biowulf PC/Linux cluster at the National Institutes of Health, Bethesda, MD (http://biowulf.nih.gov). The authors thank the reviewers for their helpful comments.

REFERENCES

1. Satagopan JM, Panageas KS. A statistical perspective on gene expression data analysis. *Statistics in Medicine* 2003; **22**:481–499.
2. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 1995; **57**:289–300.
3. Korn EL, Troendle JF, McShane LM, Simon R. Controlling the number of false discoveries: application to high-dimensional genomic data. *Journal of Statistical Planning and Inference* 2004; **124**:379–398.
4. Ge Y, Dudoit S, Speed TP. Resampling-based multiple testing for microarray data analysis (with discussion). *TEST* 2003; **12**:1–77.
5. Li SS, Bigler J, Lampe JW, Potter JD, Feng Z. FDR-controlling testing procedures and sample size determination for microarrays. *Statistics in Medicine* 2005; **24**:2267–2280.
6. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *Journal of Computational Biology* 2000; **6**:819–837.
7. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* 2001; **8**:37–52.
8. Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* 2001; **96**:1151–1160.
9. Kendziorski CM, Newton MA, Lan H, Gould MN. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Statistics in Medicine* 2003; **22**:3899–3914.
10. Kooperberg C, Aragaki A, Strand AD, Olson JM. Significance testing for small microarray experiments. *Statistics in Medicine* 2005; **24**:2281–2298.
11. Westfall PH, Young SS. *Resampling-Based Multiple Testing*. Wiley: New York, 1993.
12. Dudoit S, Yang YH, Callow MH, Speed TP. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 2002; **12**:111–139.
13. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences* 2001; **98**:5116–5121.
14. BRB-ArrayTools. http://linus.nci.nih.gov/BRB-ArrayTools.html

15. Chu G, Narasimhan B, Tibshirani R, Tusher V. *SAM* '*Significance Analysis of Microarray*' *Users Guide and Technical Document*. http://www-stat.stanford.edu/~tibs/SAM, 2005.
16. Storey JD, Tibshirani R. SAM thresholding and false discovery rates for detecting differential gene expression in DNA microarrays. In *The Analysis of Gene Expression Data*, Parmigiani G, Garrett ES, Irizarry RA, Zeger SL (eds). Springer: New York, 2003; 272–290.
17. Storey JD. Discussion to paper of Ge *et al. TEST* 2003; **12**:52–60.
18. Wright GW, Simon RM. A random variance model for detection of differential gene expression in small microarray experiments. *Bioinformatics* 2003; **19**:2448–2455.
19. Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, Meltzer P, Gusterson B, Esteller M, Kallioniemi O, Wilfond B, Borg A, Trent J. Gene expression profiles of hereditary breast cancer. *New England Journal of Medicine* 2001; **344**:539–548.
20. Simon RM, Korn EL, McShane LM, Radmacher MD, Wright GW, Zhao Y. *Design and Analysis of DNA Microarray Investigations*. Springer: New York, 2003; 168.
21. Lee MT, Gray RJ, Bjorkbacka H, Freeman MW. Generalized rank tests for replicated microarray data. *Statistical Applications in Genetics and Molecular Biology* 2005; **4**(1):article 3.
22. Jung S-H, Owzar K, George SL. A multiple testing procedure to associate gene expression with survival. *Statistics in Medicine* 2005; **24**:3077–3088.